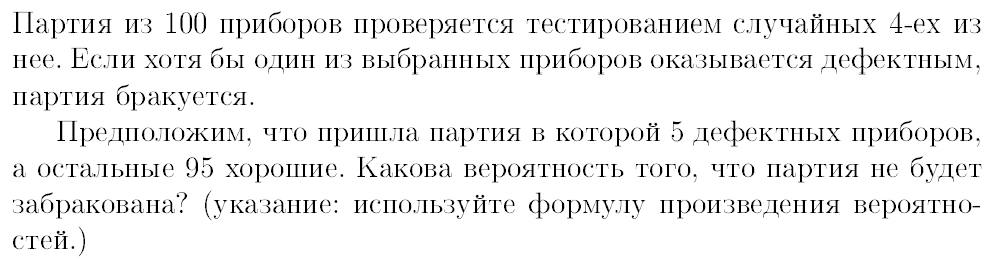
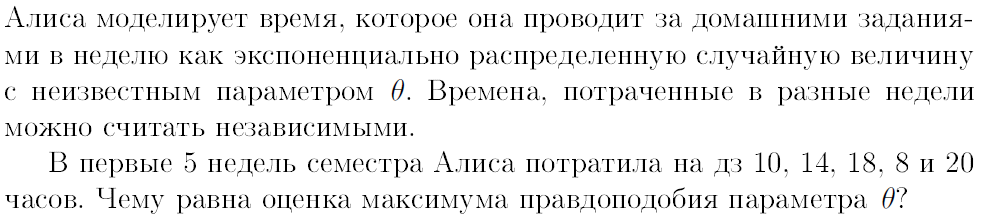
**Задача 1.**



**Задача 2.**



**Задача 3.**

Имеется информация по клиентам (выборка для разработки appl\_score\_sample.csv):

|  |  |
| --- | --- |
| CLIENT\_ID | ID клиента |
| SEX | пол (0 - женский, 1 - мужской) |
| CHILD\_FLAG | признак наличия детей (1/0-есть/нет) |
| REALTY\_FLAG | признак наличия недвижимости (1/0-есть/нет) |
| ACCOUNTS\_FLAG | признак счета в Банке (1/0-есть/нет) |
| E\_MAIL\_FLAG | признак наличия электронной почты (1/0-есть/нет) |
| DCI | доход в руб. |
| PROFIT\_FAMILY | доход семьи в руб. |
| UCI | доход по неосновному месту работы в руб. |
| FOREIGN\_PASSPORT\_FLAG | признак наличия паспорта (1/0-есть/нет) |
| STANDING\_IN\_MONTHS\_LAST | стаж работы на последнем месте работы в мес. |
| EDUCATION | образование |
| TRANSPORT\_AMOUNT | количество ТС в собственности |
| Target | целевой признак (который нужно предсказать) |

|  |
| --- |
| EDUCATION |
| 1 - ученая степень/MBA |
| 2 - второе высшее |
| 3 - высшее; |
| 4 - неоконченное высшее; |
| 5 - - среднее специальное |
| 6 - среднее; |
| 7 - ниже среднего; |
| -1 - не указано |

Необходимо любым способом (на выбор кандидата) с наибольшей точностью с помощью этой информации спрогнозировать целевой признак Тarget (бинарный признак 1/0).

1. Выбор метода моделирования и итоговой модели обосновать. Обоснование необходимо сделать по максимально возможному количеству критериев.
2. Является ли Ваш прогноз стабильным? Не переобучили ли Вы модель? (обосновать)

**Задача 4.**

Написать код SQL – запроса к таблице данных из задачи 3 (каждая задача решается одним запросом):

1. Количество уникальных/неуникальных Client\_id.
2. Посчитать долю значений целевого признака (target=1) в разрезе категорий образования.
3. На сколько в среднем доход семьи заемщиков с Client\_id=2n, больше дохода семьи заемщиков с Client\_id=2n+1, где n - произвольное целое число.

**Задача 5.**

1. В dataframe с именем subset\_1 сохраните только первые 20 строк и только 1 и 5 колонку.
2. В dataframe subset\_2 сохраните только тех клиентов-мужчин, у которых значения переменной STANDING\_IN\_MONTHS\_LAST строго больше 50.
3. Создайте две новые переменных:

CI = DCI + UCI

CI\_ln = натуральный логарифм переменной CI.

1. Для данных рассчитайте среднее значение переменной CI для каждого типа образования (переменная EDUCATION), в получившемся dataframe переменная EDUCATION не должна превратиться в индекс. Также переименуйте колонку со средним значением CI в MEAN\_CI.
2. В переменной TRANSPORT\_AMOUNT замените все отрицательные значения на медианное значение переменной TRANSPORT\_AMOUNT (без учета отрицательных значений, разумеется).